基于深度学习的广告布局图片美学属性评价

沈佳敏,鲍秉坤
(南京邮电大学 通信与信息工程学院,江苏 南京 210000)

摘 要:图像质量美学评价是近十年来比较热门的课题,但是研究的大多是对自然图像的美学评价。然而随着互联网技术的发展,线上广告业务得到了迅速发展,因此准确高效地评价一张广告布局图片的好坏是很有必要的。所谓广告布局图片,即广告图片不考虑广告语的具体内容。为了研究广告布局图片的质量美学评价,引入了一个新的数据集ALID,该数据集包含了四个美学属性的数值评分和语言评价;提出了美学多属性网络,该网络包含了三个部分:多属性特征网络、注意网络和语言生成网络。多属性特征网络通过4个不同的美学属性得分的多任务回归计算不同属性的特征矩阵,注意网络动态地调整所获特征的维度,最后语言生成网络通过长短期记忆网络生成图像字幕。实验结果表明,根据图像字幕的评价标准,该文设计的模型优于传统的CNN-LSTM模型和现代的SCA-CNN模型。

关键词:广告布局图片;美学评价;美学属性标题和美学得分;深度学习;多任务学习
中图分类号:TP181 **文献标识码**:A **文章编号**:1673-629X(2021)03-0039-06
doi:10.3969/j.issn.1673-629X.2021.03.007

Aesthetic Attribute Evaluation of Advertising Layout Images Based on Deep Learning

SHEN Jia-min, BAO Bing-kun
(School of Communication and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210000, China)

Abstract: The aesthetic evaluation of image quality is a hot topic in the past decade, but most of the research is about the aesthetic evaluation of natural images. However, with the development of Internet technology, online advertising business has been developed rapidly, so it is necessary to accurately and efficiently evaluate the quality of an advertising layout images. The so-called advertisement layout images mean that the advertisement images do not take into account the specific content of the advertisement language. In order to study the aesthetic quality evaluation of advertising layout images, a new data set, ALID, is introduced, which includes the numerical evaluation and language evaluation of four aesthetic attributes. An aesthetic multi-attribute network is proposed, which consists of three parts: multi-attribute feature network, attention network and language generation network. The multi-attribute feature network calculates the feature matrix of different attributes through multi-task regression of four different aesthetic attribute scores. Attention network dynamically adjusts the dimensions of acquired features. Finally, the language generation network generates image subtitles through the long short-term memory. The experiment shows that the proposed model is superior to the traditional CNN-LSTM model and the modern SCA-CNN model according to the evaluation criteria of image subtitles.

Key words: advertisement layout images; aesthetic evaluation; aesthetic attribute title and aesthetic score; deep learning; multi-task learning

0 引 言

随着互联网技术的发展,广告业务有了新的展现形式:在线广告,而线上线下服务的打通,使得在线广告的需求越来越大。因此,准确高效地评价一张广告图片是很有必要的。为了简化工作,文中并不关注广告图片中广告语的具体内容,即文中对广告布局图片进行美学属性评价。

对于一个人类艺术家来说,当他/她展示一张照片或一幅图画时,他/她不仅会从不同的美学属性方面给出一个数字分数,而且还会说出一个段落来描述不同

收稿日期:2020-04-22 **修回日期**:2020-08-25
基金项目:国家自然科学基金面上项目(61772287);江苏省高等学校自然科学研究重大项目(18KJA510004)
作者简介:沈佳敏(1995-),男,硕士研究生,研究方向为计算机视觉;鲍秉坤,博士,教授,CCF会员(34665M),研究方向为人工智能、计算机视觉、多媒体计算、社交多媒体。

的美学属性[1]。一张图片的美学属性包括构图、灯光、颜色、图片的焦点等。在这项工作中,对于广告布局图片的评分标准[2],该文主要考虑构图、色彩照明、图片的焦点以及总体印象。

构图主要考虑的是整张图片的稳定性[3]。所谓稳定,是人类在长期观察中自然形成的一种视觉习惯和审美观念。因此,凡符合这种审美观念的造型艺术才能产生美感,违背这个原则的,看起来就不舒服。但是,稳定不意味着图片的元素在图片中平均分配,而是所有的元素在图片中存在一种合乎逻辑的比例关系,例如对称分布。事实上,对称的稳定感特别强,对称能使图片有庄严、肃穆、和谐的感觉,像中国古代的建筑就是对称的典范。

颜色是通过眼、脑和生活经验所产生的对光的视觉感受[4],肉眼所见到的光线,是由波长范围很窄的电磁波产生的,不同波长的电磁波表现为不同的颜色,对色彩的辨认是肉眼受到电磁波辐射能刺激后所引起的视觉神经感觉。颜色具有三个特性,即色相、明度和饱和度。

图片的焦点是图片中的重要组成部分[5],如果没有固定的兴趣点,图片就会显得杂乱无章。没有什么能吸引观众的注意力或引起他们的兴趣,也没有什么线索能说明图片的目的是什么。但另一方面,具有强烈兴趣点的照片会立即向观众展示照片的全部内容。它们引起了人们的注意,并把观众吸引到一个构图中,让他们的眼睛停留片刻。在这项工作中,广告图片的焦点即是突出广告的主体。

图片的总体印象,即对整张图片的感受,主要考虑整张图片各个组成部分放在一起是否适宜,以及组合到一起后整体的美学印象。

图像字幕,大多数图像字幕工作遵循 CNN-RNN 框架,取得了很好的效果[6]。近年来关于图像字幕[7-8]的文献大多介绍了注意方案[9]。该文遵循这一趋势,在网络中添加注意力模式。

为了得到广告布局图片有关上述四个方面的美学属性评价,文中设计了美学多属性网络,包括多属性特征网络、注意网络以及语言生成网络,然后根据图像字幕的评价标准,比较文中模型和其他模型。

1 数据集

由于没有适合该文场景的公开数据集,因此选择 5 名专业的广告设计师以及 10 名广告从业人员对广告布局图片进行了美学属性评价,数据集中包含各美学属性的数值评分和语言评价。该文将这个数据集称为 ALID 数据集,该数据集的美学属性包括色彩照明、构图、景深、焦点和总体印象。数据集中总共有大约 200 000张广告布局图片,对于每个属性,选择 2 000 张图片进行验证,2 000 张图片进行测试,剩下的图片用于训练。

2 系统模型

在本节中将详细介绍整个模型的系统框架。如图 1 所示,所提出的美学多属性网络(AMAN)分为三个部分:多属性特征网络(multi-attribute feature network,MAFN)、注意网络和语言生成网络(language generation network,LGN)。MAFN 通过 4 个属性得分的多任务回归计算不同属性的特征矩阵。注意网络动态地调整所获得特征的通道维度和空间维度的注意权重。最后,LGN 通过长短期记忆网络(long short-term memory,LSTM)[10]生成字幕,LSTM 网络需要数据集中关于美学语言评价的真实内容和注意网络调整后的特征映射。

2.1 多属性特征网络

多任务学习是训练深卷积网络的常用方法。由于属性的多样性,多任务学习可以通过参数共享实现美学的多属性评价。审美属性评价相对独立。然而,模型训练过程是相似的。在 ALID 中,除了每个属性的分数外,每个图像都有一个全局分数。因此,MAFN 的损失分为两部分。一个是每个属性(m 个属性)的损失。另一个是全局损失。N 表示批处理中的图片数,$\hat{y}^i$ 表示网络最后一个完全连接层的输出,y^i 表示真分数。公式如下表示:

$$\mathrm{Loss}_{\mathrm{attribute}} = \mathrm{Loss}_{\mathrm{global}} = \frac{1}{2N}\sum_{i=1}^{N} \| \hat{y}^i - y^i \|_2^2 \tag{1}$$

$$\mathrm{Loss} = \sum_{i=1}^{m} \mathrm{Loss}_{\mathrm{attribute}} + \mathrm{Loss}_{\mathrm{global}} \tag{2}$$

MAFN 如图 1 上部所示,GFN 和 AFN 使用 Desnet161 来提取密集的特征图。所有先前图层的参数都是共享的。GFN 和 AFN 的输出分为 5 个部分:一般特征和 4 个审美属性的特征。GFN 对全局美学分数的输出执行全连接操作。

对于最终结果,执行均方误差(MSE)[4]的计算,并将其作为模型损失参数返回到之前的层。AFN 对属性特征映射进行卷积运算,得到 4 个不同的属性特征映射。与 GFN 一样,通过全连接层和均方误差损失得到最终属性得分。

MAFN 可以同时提取图像的不同属性特征映射。因此,该模型不再局限于输出一个句子的注释。图像的美学特征可以从多个属性进行评价,更好地指导图像的综合评价。多任务网络得到的具体结果还可以直接利用知识迁移扩展 ALID 字幕数据集的属性评估,从而提供更广泛的审美评估能力。

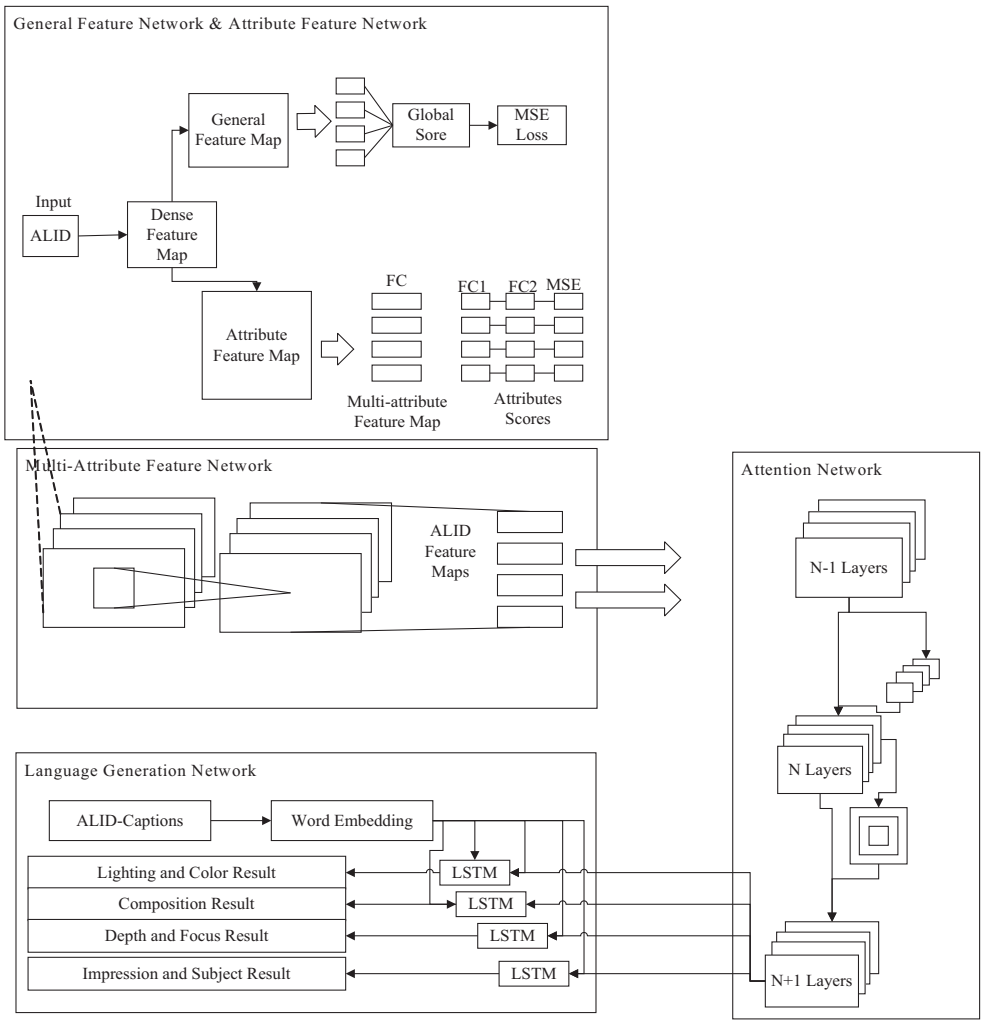

图 1　系统框图

2.2　注意力网络

注意力网络包含两种模式[11],一是空间注意在信道注意之后,二是信道注意在空间注意之后。通过实验,该文采用第一种结构作为注意力网络。给定特定的 $N-1$ 层特征映射 M_{N-1},根据信道注意计算 f_c,得到信道注意权重 w_c。然后将加权 w_c 和 $N-1$ 层特征映射线性融合,得到新的 N 层信道感知特征映射 M_N。之后,将信道感知特征图 M_N 发送到空间感知注意模块计算 f_s,得到了空间注意权重 w_s。最后,对前一步得到的信道感知特征图 M_N 进行空间感知,即 CNN 输出的特征。合并的过程可以用下面的公式表示:

$$f_c = \tanh((w_c \otimes M_{N-1} + b_c) \oplus w_{hc} h_{t-1}) \tag{3}$$

$$M_N = \mathrm{softmax}(W_N f_c + b_N) \tag{4}$$

$$f_s = \tanh((w_s \otimes M_{N-1} + b_s) \oplus w_{hs} h_{t-1}) \tag{5}$$

$$M_{N+1} = \mathrm{softmax}(W_N f_s + b_N) \tag{6}$$

在上面的公式中,t 表示时间状态,h 表示 LSTM 隐藏状态,$\oplus$ 表示矩阵和向量相加,$\otimes$ 表示向量的外积,b 表示偏移量。

2.3　语言生成网络

长短期记忆网络是学习长期依赖信息的一种特殊类型的 RNN。在许多问题上,LSTM 已经取得了相当大的成功,并得到了广泛的应用。通过将多个属性的信息输入 LSTM 单元,可以根据图像特征和时序信息进行下一个单词的预测。具体来说,如果美学评估的两个子任务和生成的注释是统一的,则训练过程可以描述为这样的形式:对于训练集的图片 I,对应的描述是序列 $S = \{S_1, S_2, \cdots, S_N\}$(其中 S_i 表示句子)。对于语言生成模型 θ 和属性 ∂,给定输入图片 I,为每个属性生成序列 S_i 的概率如下:

$$P_\partial(S \mid I) = \prod_{t=0}^{N} P_\partial(S_t \mid S_0, S_1, \cdots, S_{t-1}, I; \theta_\partial) \tag{7}$$

该模型利用通道和空间注意模型来提高图像有效区域的利用率。因此,可以在解码阶段更有效地利用图像的特定区域的特征。语言生成网络的损失可以用下面的公式来表示:

$$\mathrm{Loss}_\partial(I, S) = -\sum_{t=1}^{N} \log P_t(S_t) \tag{8}$$

该模型利用图像的语义信息来指导解码阶段的词序生成,避免了仅在解码开始时使用图像信息的问题,从而导致图像信息随着时间的推移逐渐丢失。为了更

好地获取图像的高层语义信息,该模型对原有的卷积神经网络进行了改进,包括多任务学习方法,该方法可以提取图像的高层语义信息,增强编码阶段图像特征的提取。

3 实验结果展示及分析

3.1 基准网络

CNN – LSTM:该模型基于 Google 的 NIC 模型[12]。Resnet–152[13]提取不同属性的特征,LSTM 进行编码。该基线与文中方法的区别在于:(1)没有引入注意机制来增强特征提取过程;(2)没有使用多任务网络来提取不同属性的特征。相反,每个属性分别训练一个网络。它没有充分利用 CNNs 的美学特征,在提取 CNNs 特征时会进行简单的知识转移。

SCA – Model:该模型基于 SCA – CNN[12]模型,ResNet–152 为不同的属性提取特征。LSTM 在提取特征后进行空间和通道注意增强。此基线与文中基线的区别在于:(1)SCA 模型不使用多任务网络来提取不同属性的特征。每个属性分别训练一个网络;(2)SCA 模型没有充分利用美学特征。在提取 CNNs 的特征时,会发生一个简单的知识迁移。

3.2 实验细节

文中实验基于该框架,LSTM 单元数为 1 000 个,发送到 LSTM 单元的特征包括 2 048 维全局特征和 512 维属性特征。单词向量维数设置为 50。基础学习率为 0.01。注意模块的尺寸为 512。在训练的过程中采用 dropout,以防止过度拟合。采用随机梯度下降优化策略对网络进行优化。

3.3 结果展示

模型的测试结果如图 2 所示。可以发现,结果不仅具有丰富的句子结构,而且对特征的把握也非常准确。评论和属性的相关性很高。在得分方面,平均属性得分非常接近地面真相得分。通过评分和点评,形象评价生动。

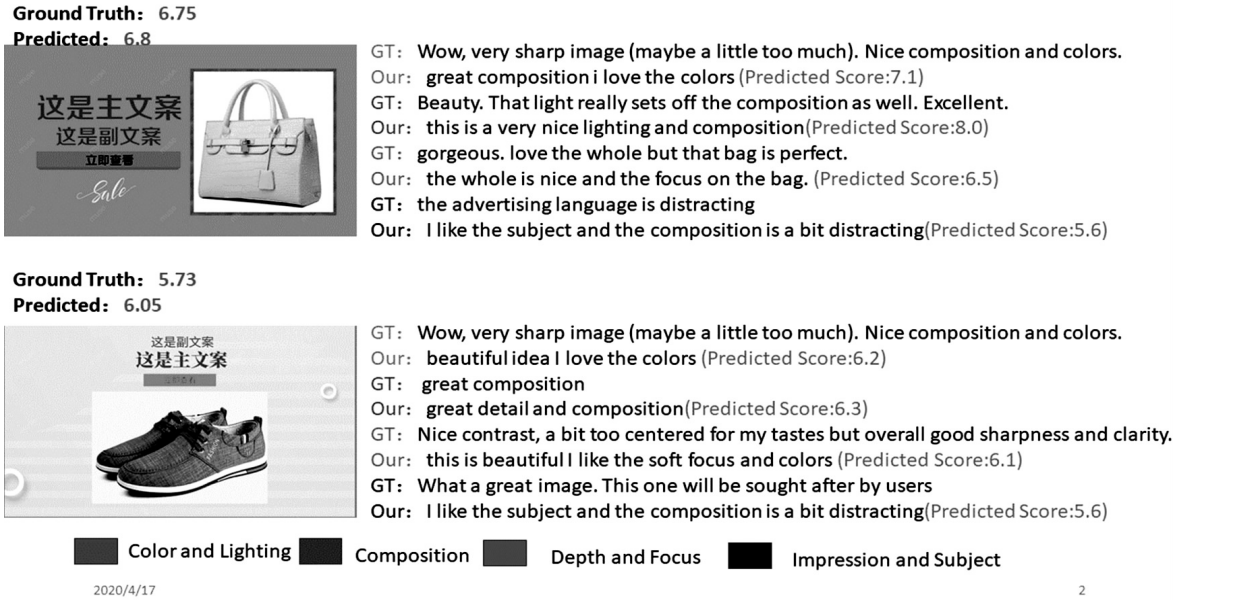

图 2 结果展示

3.4 比较分析

3.4.1 与基准网络比较

比较文中模型和基准网络,性能的评估标准包括 RLEU–1、2、3、4、METEOR、ROUGE 和 CIDEr。表 1 和表 2 所示的比较结果表明,文中模型在所有标准中都优于基线模型。

表 1 文中模型和基准网络关于 BELU 准则的比较结果

Method	BELU–1	BELU–2	BELU–3	BELU–4
CNN–LSTM(Color and Lighting)	46.3	23.2	13.6	6.9
SCA–Model(Color and Lighting)	46.5	23.3	13.9	7.1
AMAN(Color and Lighting)	48.7	25.0	14.4	7.3
CNN–LSTM(Composition)	47.5	24.5	14.1	7.0
SCA–Model(Composition)	48.0	24.6	14.3	7.2
AMAN(Composition)	49.2	24.9	14.6	7.5

续表 1

Method	BELU-1	BELU-2	BELU-3	BELU-4
CNN-LSTM(Depth and Focus)	46.2	23.1	13.2	6.4
SCA-Model(Depth and Focus)	46.3	23.3	13.4	6.5
AMAN(Depth and Focus)	47.1	24.0	13.7	6.8
CNN-LSTM(Impression and Subject)	46.1	23.0	13.5	6.9
SCA-Model(Impression and Subject)	46.2	23.3	13.5	7.0
AMAN(Impression and Subject)	46.8	23.6	13.9	7.4

表 2　文中模型和基准网络关于其他准则的比较结果

Method	METEOR	ROUGE	CIDEr
CNN-LSTM(Color and Lighting)	12.5	27.0	6.1
SCA-Model(Color and Lighting)	12.8	27.4	6.2
AMAN(Color and Lighting)	13.2	27.9	6.5
CNN-LSTM(Composition)	12.7	28.2	6.4
SCA-Model(Composition)	12.8	28.6	6.5
AMAN(Composition)	13.6	28.9	6.8
CNN-LSTM(Depth and Focus)	12.2	26.8	6.0
SCA-Model(Depth and Focus)	12.3	26.9	6.0
AMAN(Depth and Focus)	12.7	27.7	6.3
CNN-LSTM(Impression and Subject)	12.6	27.2	6.1
SCA-Model(Impression and Subject)	12.6	27.3	6.2
AMAN(Impression and Subject)	13.0	27.6	6.7

3.4.2　与其他方法比较

使用 SPICE[14] 来比较方法[1]和文中模型之间的性能。SPICE 是自动评估生成的图像标题的标准。它通过将句子解析成一个图来解决结果和生成的标题之间的相似性。计算公式如下:

$$\mathrm{SPICE} = \mathrm{F1Score} = \frac{2 * \mathrm{Precision} * \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}} \qquad (9)$$

如表3 所示,该模型在各种属性上都优于文献[9]提出的方法。方法[1]采用属性融合训练方法,将构图、色彩和光照、主题三个属性结合起来。但是,通过对比可以发现,文中在这三个属性中生成的注释比前面的注释具有更好的注释。

表 3　结果在测试集上通过 SPICE 准则的比较

Method	SPICE	Precision	Recall
CNN-LSTM-WD[1]	0.136	0.181	0.156
AO Approach[1]	0.127	0.201	0.121
AF Approach[1]	0.150	0.212	0.157
CNN-LSTM(Color and Lighting)	0.166	0.179	0.154
SCA-Model(Color and Lighting)	0.174	0.194	0.158
AMAN(Color and Lighting)	0.196	0.231	0.170
CNN-LSTM(Composition)	0.167	0.184	0.153
SCA-Model(Composition)	0.178	0.203	0.159
AMAN(Composition)	0.197	0.228	0.174
CNN-LSTM(Depth and Focus)	0.163	0.174	0.153
SCA-Model(Depth and Focus)	0.167	0.182	0.154
AMAN(Depth and Focus)	0.187	0.215	0.165
CNN-LSTM(Impression and Subject)	0.158	0.169	0.149
SCA-Model(Impression and Subject)	0.162	0.175	0.150
AMAN(Impression and Subject)	0.181	0.213	0.158

3.5 结 论

提出了一个新的问题:广告布局图片的美学质量评价,建立了一个新的数据集 ALID,提出了一种新的网络 AMAN,该网络可以生成美学标题和美学属性数值评分。

4 结束语

图像美学质量评价是比较热门的研究问题,该文研究了广告布局图片的美学质量属性评价。为了研究这个问题,邀请了 5 名专业的广告设计师以及 10 名广告从业人员对广告布局图片进行美学属性评价,从而构造了包含各美学属性的数值评分和语言评价的新的数据集 ALID,提出了美学评价模型 AMAN。该模型包含了多属性特征网络(MAFN)、注意网络和语言生成网络(LGN)。通过实验分析,该模型在各个评价标准下都表现得比较优异。当然,该方法仍有需要完善的地方,例如可以考虑利用强化学习来生成语言评价。

参考文献:

[1] CHANGK Y,LU K H,CHEN C S. Aesthetic critiques generation for photos[C]//2017 IEEE international conference on computer vision (ICCV). Venice,Italy:IEEE,2017:3534-3543.

[2] 王文涓. 广告与后现代美学[J]. 太原师范学院学报:社会科学版,2011,10(3):26-28.

[3] 王钟涛. 灵活多变的构图法则[N]. 美术报,2020-01-25(023).

[4] 王洪姣. 图像质量评价方法的研究及实现[D]. 西安:西安电子科技大学,2014.

[5] 陈 杰. 一种红外图像中基于焦点选择的目标检测识别方法[J]. 科技传播,2018,10(7):102-103.

[6] DONAHUE J,HENDRICKS L A,GUADARRAMA S,et al. Long-term recurrent convolutional networks for visual recognition and description[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2017,39(4):677-691.

[7] 刘 恒. 协同常规-特定语义的多特征图像字幕生成[D]. 西安:西安电子科技大学,2019.

[8] 杜海骏,刘学亮. 融合约束学习的图像字幕生成方法[J]. 中国图象图形学报,2020,25(2):333-342.

[9] ANDERSON P,HE X,BUEHLER C,et al. Bottom-up and top-down attention for image captioning and visual question answering[C]//IEEE conference on computer vision and pattern recognition. Salt Lake City,Utah,USA:IEEE,2018:6077-6086.

[10] 翟社平,杨媛媛,邱 程,等. 基于注意力机制 Bi-LSTM 算法的双语文本情感分析[J]. 计算机应用与软件,2019,36(12):251-255.

[11] SCHEN L,ZHANG H,XIAO J,et al. SCA-CNN:spatial and channel-wise attention in convolutional networks for image captioning[C]//IEEE conference on computer vision & pattern recognition. Honolulu, HI, USA: IEEE, 2017: 6298-6306.

[12] VINYALS O,TOSHEV A,BENGIO S,et al. Show and tell: a neural image caption generator[C]//IEEE conference on computer vision and pattern recognition. Boston,MA,USA:IEEE,2015:3156-3164.

[13] HE K,ZHANG X,REN S,et al. Deep residual learning for image recognition[C]//IEEE conference on computer vision and pattern recognition. Las Vegas, Nevada, USA: IEEE, 2016:770-778.

[14] ANDERSON P, FERNANDO B, JOHNSON M, et al. SPICE:semantic propositional image caption evaluation[J]. Adaptive Behavior,2016,11(4):382-398.